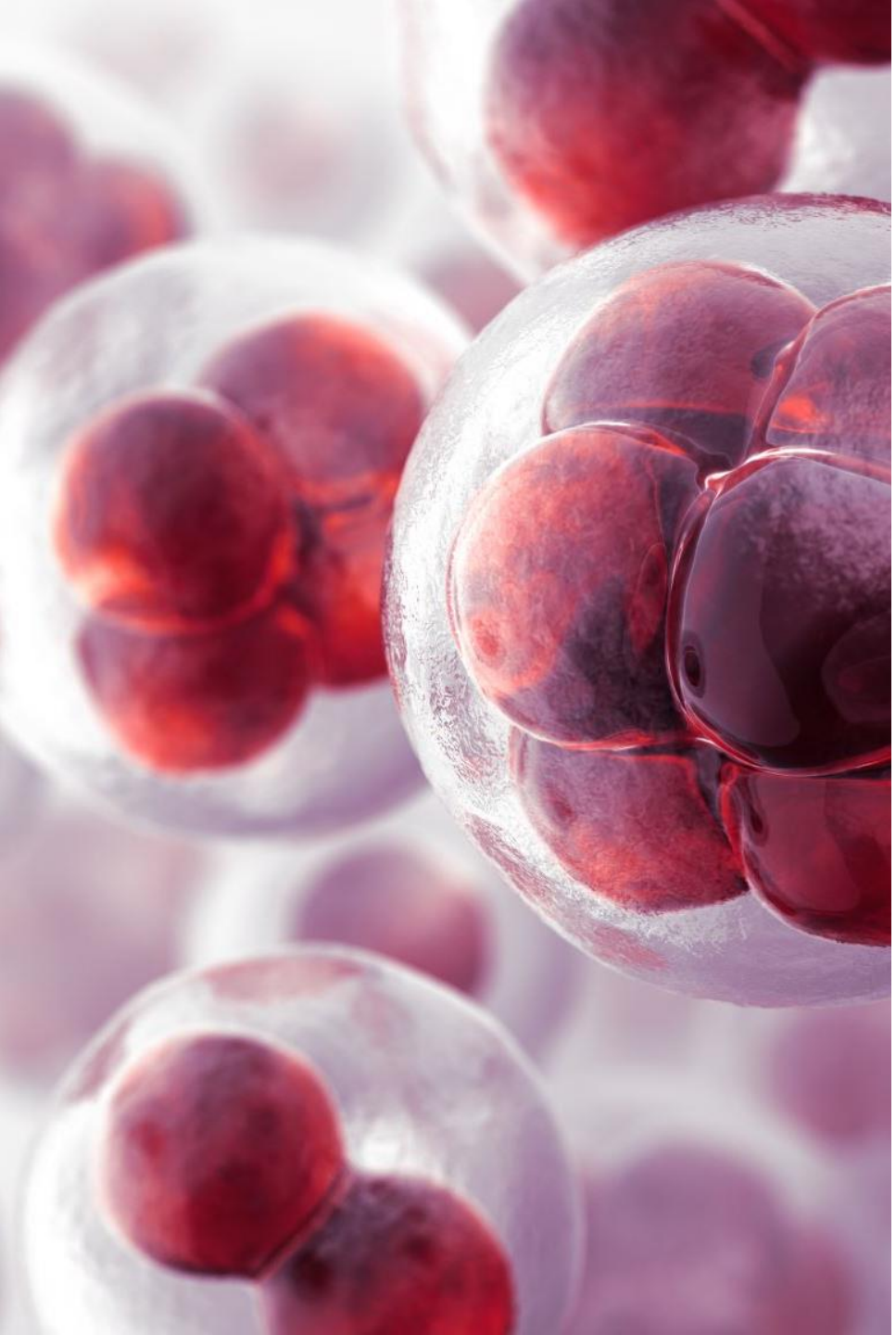
Protein domain boundary prediction based on Deep Neural Networks



What are proteins?

Proteins are essential molecules for living organisms.

Keratin is the key structural material making up hair and nails.

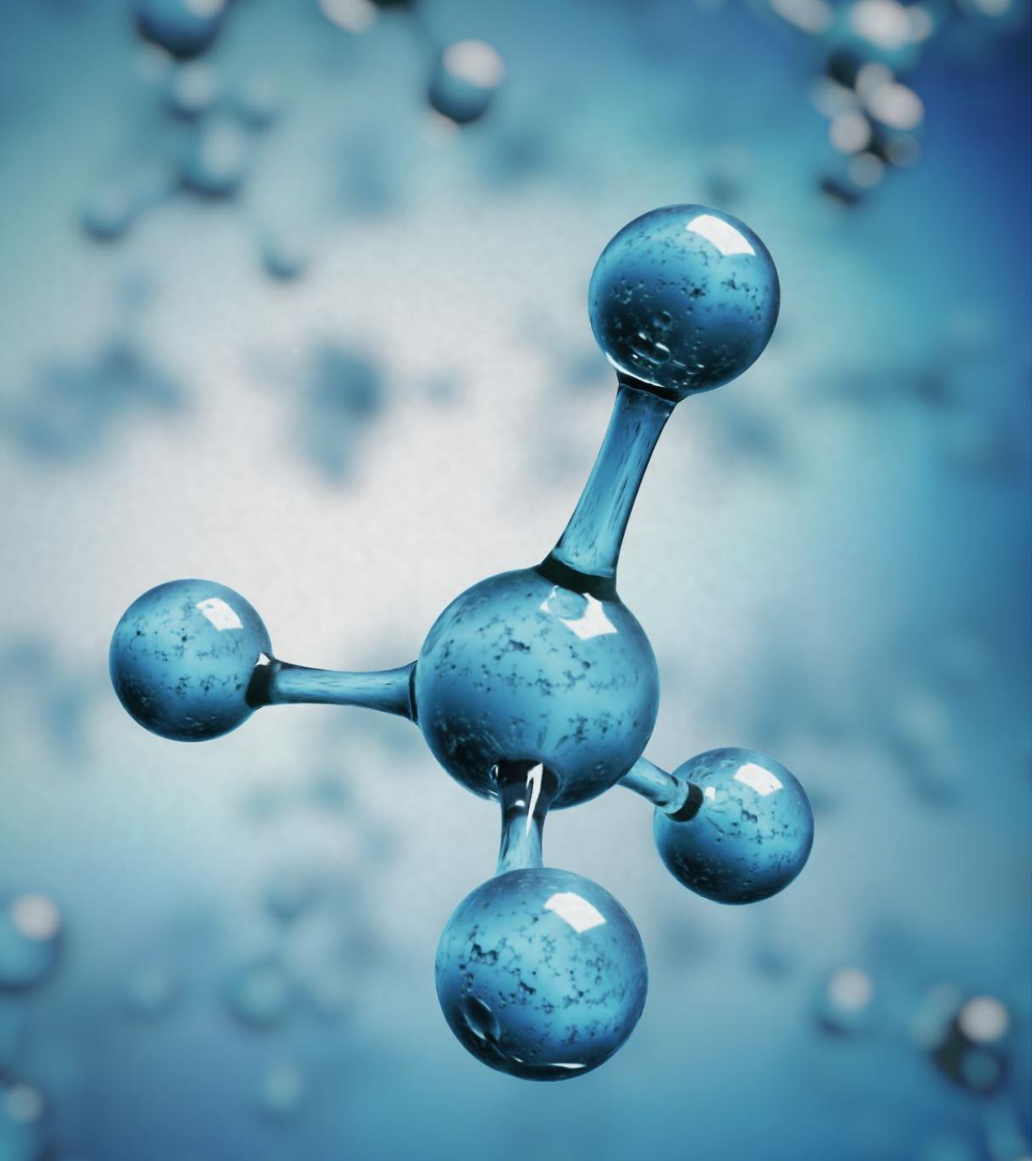
Insulin helps blood sugar enter the body cells to be used for energy.

Each protein is made up of one or more **sequences** of organic compounds called **amino acids**

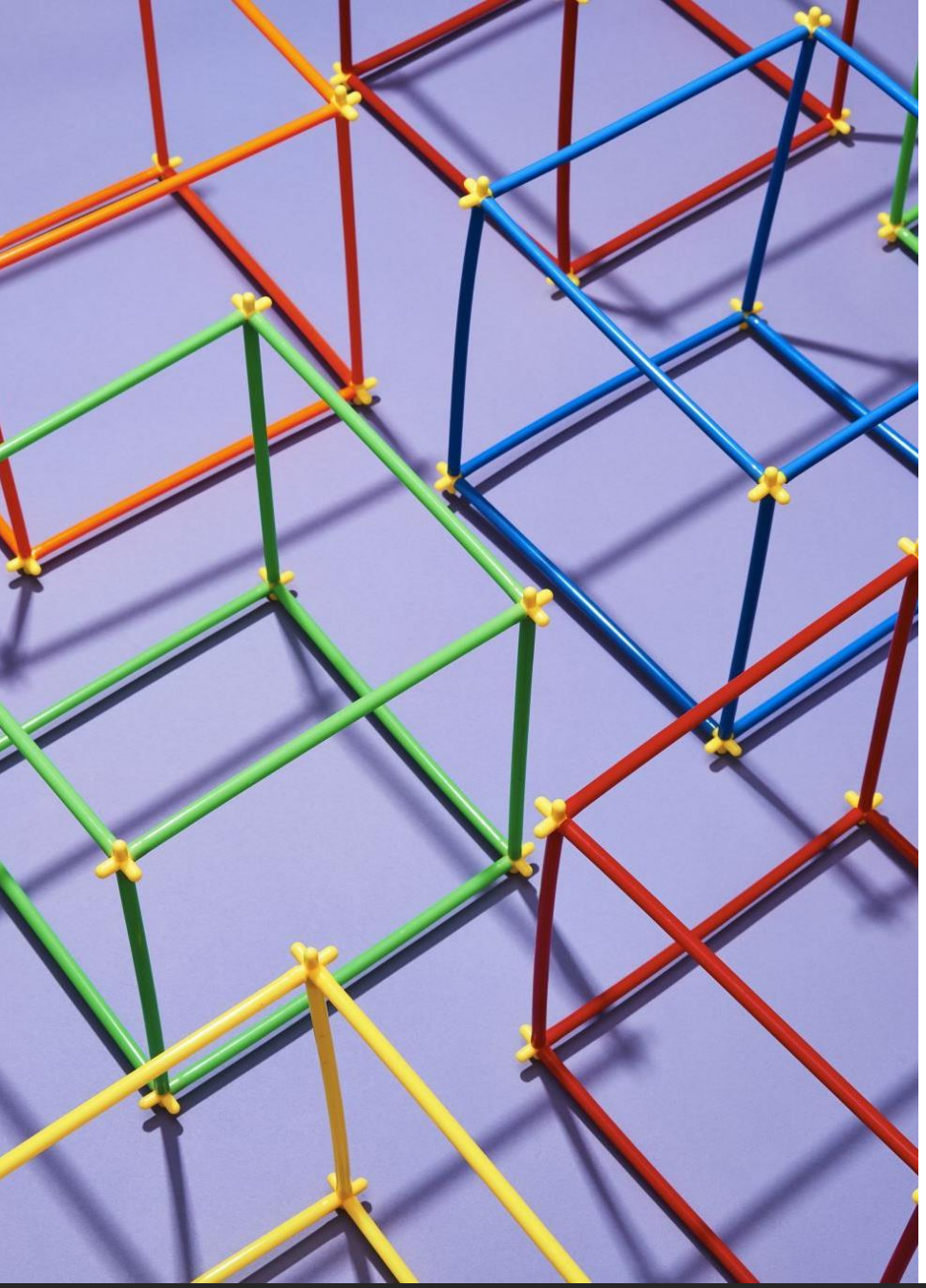
These sequences are called polypeptide chains – or **chains**.

How are proteins useful?

Drug discovery	Designing drugs to target specific proteins can aid drug discovery
Biotechnology	Engineering new proteins with desired properties can be used in biotechnology applications
Treatments	Understanding how proteins fold into a particular shape can help develop new treatments for diseases such as Parkinson's disease and Alzheimer's disease.



We must understand the physical structure of each protein first



Structure properties

We are interested in the protein sub-units that fold and function independently called **domains**

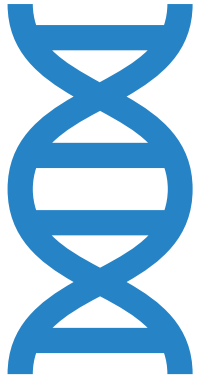
Domain boundaries separate domains

Identifying where the boundaries are **helps** identify domains and **where** these domains are in the protein

More information about the structure!

Current methods - limitations

Experimental methods such as X-ray crystallography are **expensive and time consuming**



Some proteins are **too long** or may not crystallise well

Software

Some methods based on software can predict domain boundaries **but** require the experimental 3D structure of a protein.

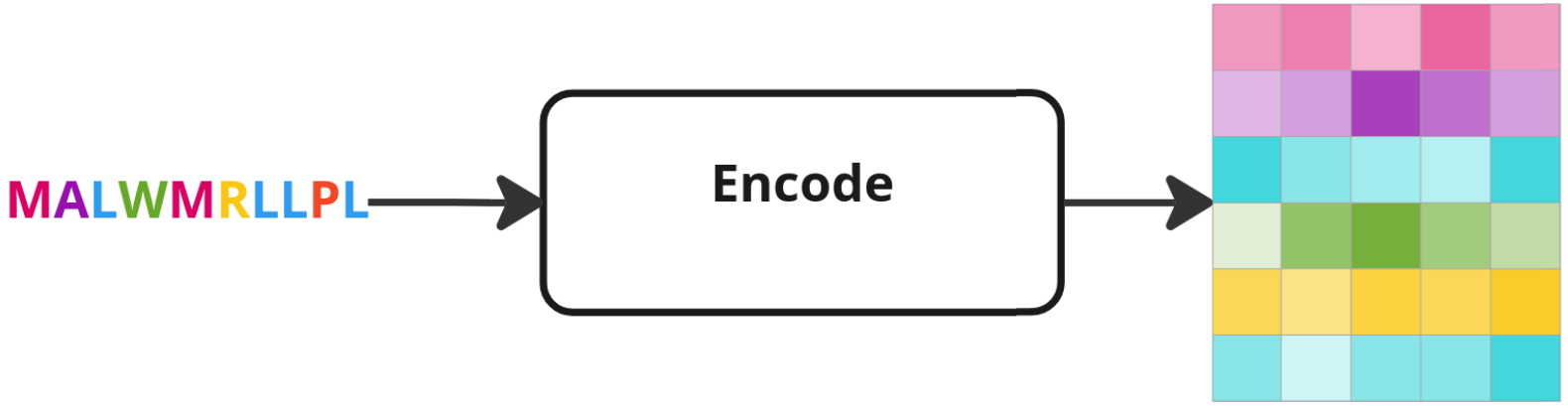
Not ideal!

Machine learning

Machine learning methods require only one input: the **amino acid sequence**

This is possible due to the Thermodynamic Hypothesis which states that *"the native structure is determined **only** by the protein's amino-acid sequence"*

How is an amino-acid sequence input to a machine learning model?



Naïve way: One-hot encoding

MALW

1 0 0 0

0 1 0 0

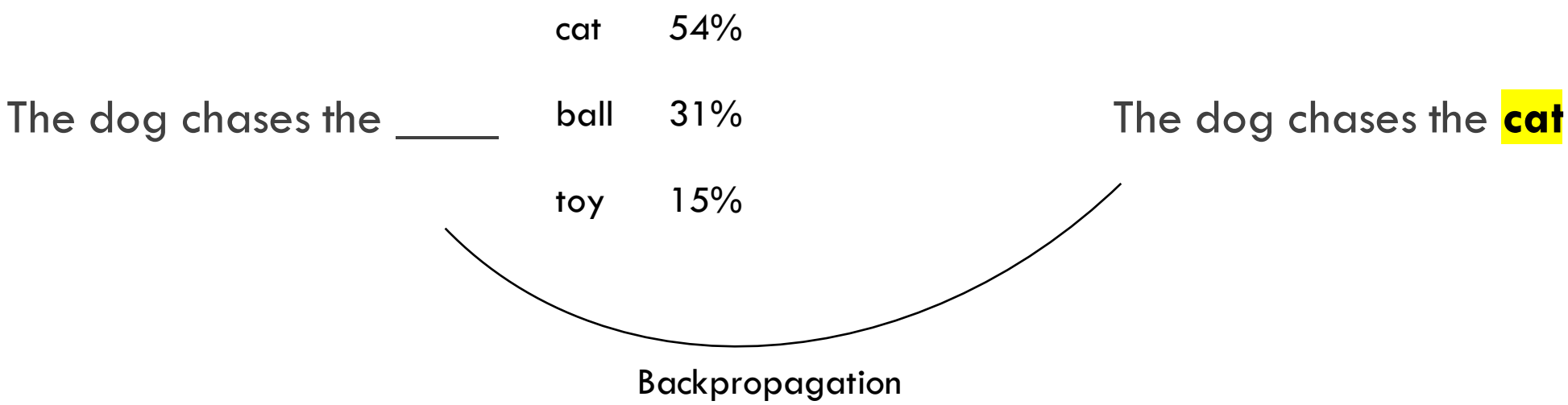
0 0 1 0

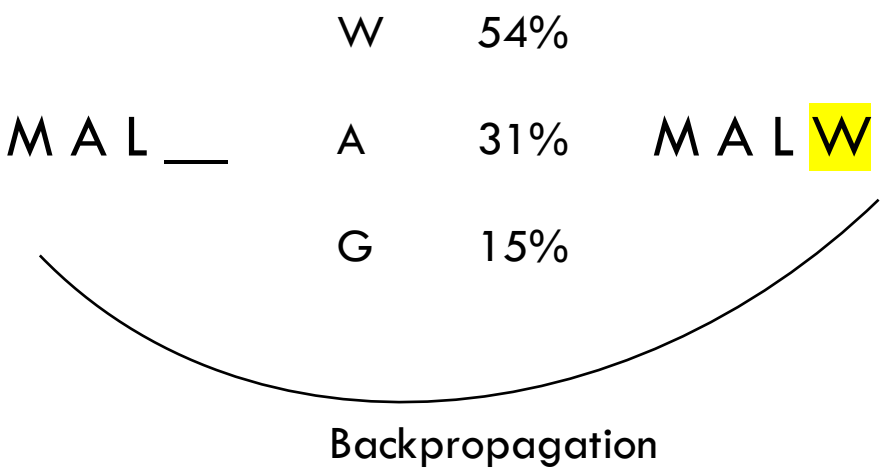
0 0 0 1

Better way: Protein language models

Similar to language models

Trained on the masked language modelling task




W 54%
M A L __
A 31%
M A L W
G 15%
Backpropagation

Protein language models learn matrix representations that capture **intrinsic information** about the amino acid sequence, which allows to predict the masked amino acids

Two state-of-the-art pre-trained protein language models. How do they compare?

Evolutionary Scale Modelling (ESM)

- Already used in the literature for domain boundary prediction
- Utilises the Transformer architecture which is very popular in natural language processing tasks such as translation
- Transformers scale quadratically with input

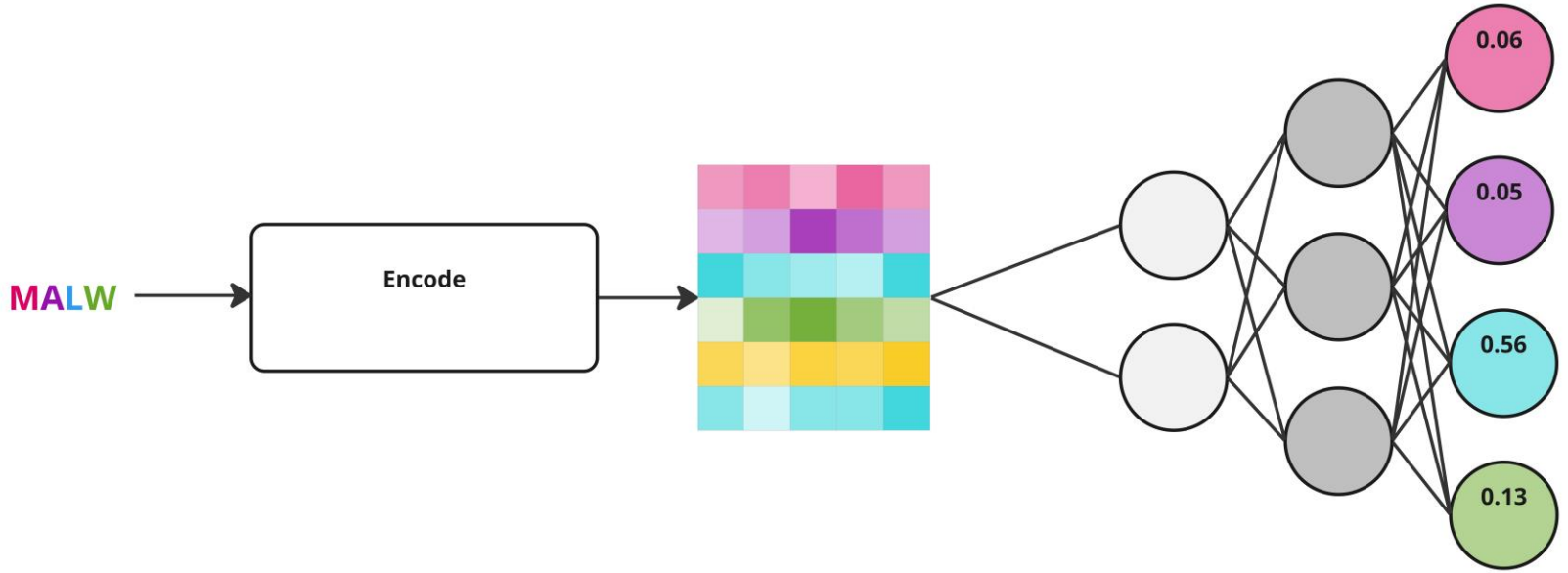
Convolutional Autonencoding Representation of Proteins (CARP)

- Has not been used in the literature yet for domain boundary prediction
- Utilises a convolutional autoencoder which scales linearly with input

How to predict domain boundaries?

We use a **deep neural network** to learn from data which amino acids in a sequence are domain boundaries

Since we are dealing with **sequential data,** we utilised a **bi-directional Long Short-Term Memory** model (LSTM)


MALW
Encode
0.06
0.05
0.56
0.13

- How do we evaluate?

Domain number prediction

Is a protein single-domain or multi-domain?

Precision, Recall, Accuracy, Mathew's Correlation Coefficient (MCC)

Boundary prediction

How close are the predicted boundaries to the true boundaries?

Domain boundary distance (DBD)

Evaluation metrics help answer our questions:

- Precision: How many predicted positives are indeed true positives?
- Recall: Out of all the true positives, how many were predicted correctly?
- Accuracy: Overall, how many times is the prediction correct?
- **MCC: How strong is the bivariate relationship between predictions and the ground truth?**
- **DBD: How close to the true boundaries are the predicted boundaries?**

Note!

For domain number metrics, the model is evaluated in classifying a protein as single-domain or multi-domain

The classification is implicit. That is, the model predicts the probability of each residue being a boundary

We use a cutoff threshold to convert probabilities into 0s or 1s

We take the sum of the predicted boundaries. The number of domains is equal to the number of boundaries **plus 1**

Why evaluate domain number prediction?

Correctly classifying a protein as single or multi-domain provides **useful information** for the protein structure

This is the **convention** used in Bioinformatics studies when evaluating protein domain boundaries

We want to **compare** our model with the state-of-the-art

Results

How do the different encoding mechanisms perform ?

Results from 5-fold-cross-validation on our dataset

Methods	Domain Number Prediction						Boundary prediction
	Single-domain		Multi-domain		All		
	Pre	Rec	Pre	Rec	Acc	MCC	DBD
ESM	**0.9244**	**0.8104**	**0.7392**	**0.9452**	**0.8494**	**0.7088**	**0.5596**
CARP	0.9009	0.7768	0.6820	0.9328	0.8175	0.6446	0.4529
One-hot	0.4643	0.0000	1.0000	0.0000	0.4643	0.0000	0.4643

Are the results statistically significant?

Results from the statistical analysis (t-tests)

Pair	MCC		DBD	
	t-statistic	p-value	t-statistic	p-value
ESM, CARP	1.6266	1.42E-01	23.4	1.6945 e-118
ESM, one-hot	33.2153	7.37E-10	68.3	0 (underflow)
CARP, one-hot	19.4431	5.09E-08	52.4	0 (underflow)

Significant results:

- Difference between ESM and one-hot for both metrics
- Difference between CARP and one-hot for both metrics
- Difference between ESM and CARP for DBD

Non-significant result:

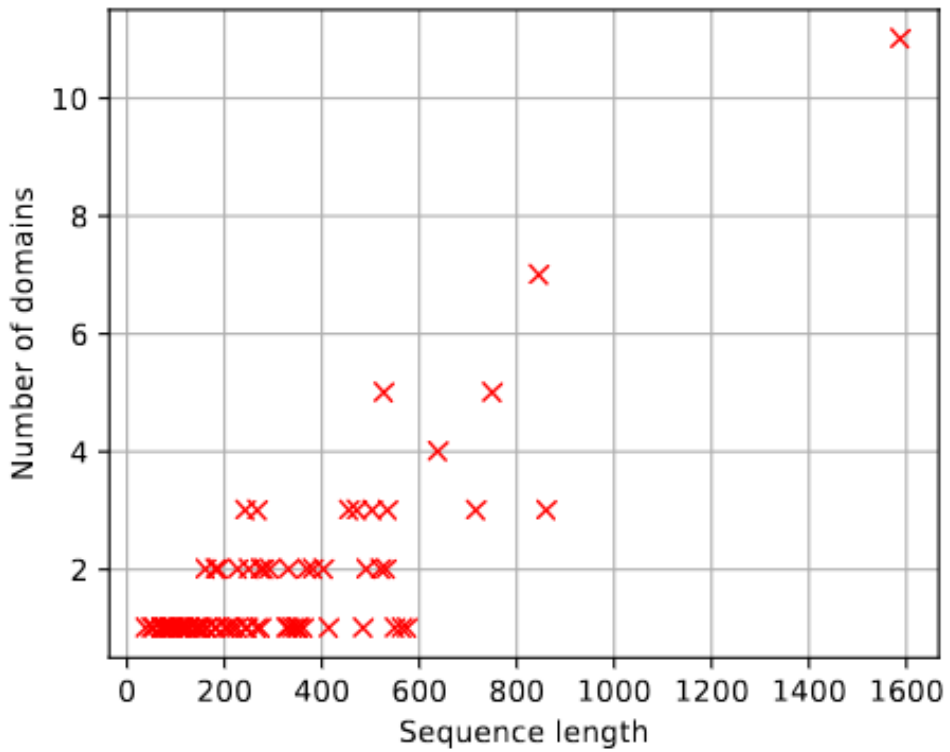
- Difference between ESM and CARP for MCC

How does our final model (ESM) perform and compare?

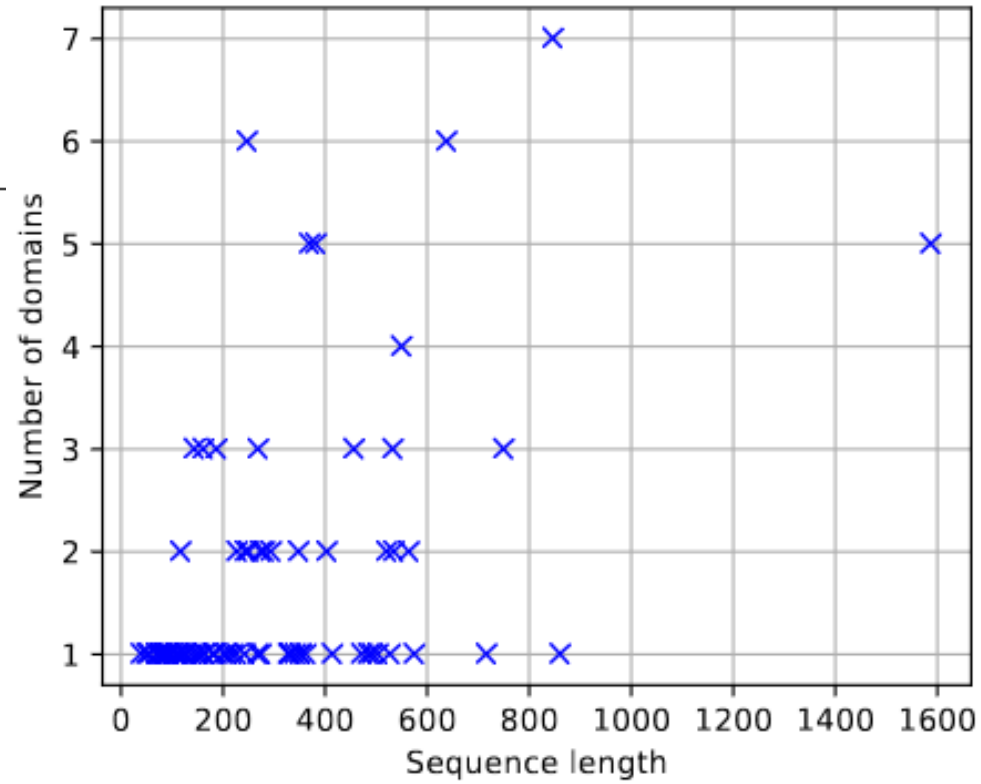
Results from testing on an independent dataset provided by CASP (Critical Assessment of protein Structure Prediction)

Methods	Domain Number Prediction						Boundary prediction
	Single-domain		Multi-domain		All		
	Pre	Rec	Pre	Rec	Acc	MCC	DBD
Res-Dom	**0.963**	**0.788**	0.667	**0.933**	**0.833**	**0.674**	**0.532**
Our model	0.865	0.679	**0.833**	0.731	0.8	0.554	0.125
FUpred	0.95	0.576	0.5	**0.933**	0.688	0.479	0.578
DNN-Dom	0.839	**0.788**	0.588	0.667	0.75	0.441	0.457

Predicted number of domains
Correlation coefficient: 0.8142



True number of domains
Correlation coefficient: 0.4908



Strong correlation between the model predicted number of domains and the sequence length suggests that the model may implicitly utilise the length during prediction.

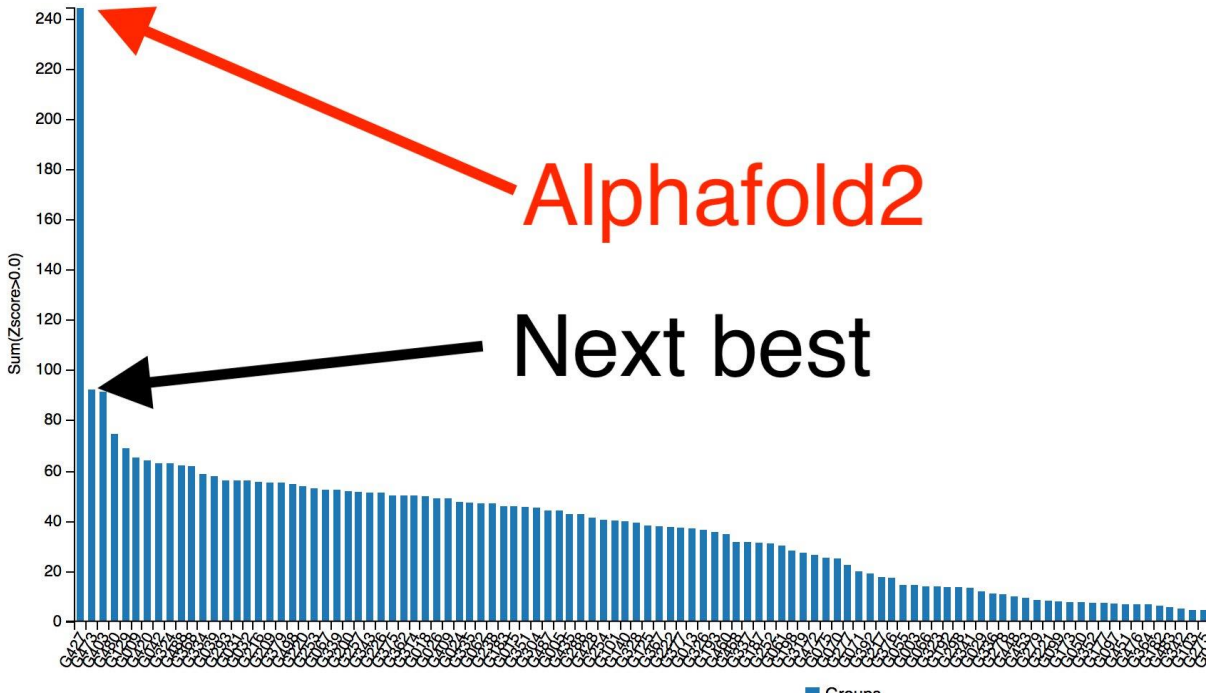
AlphaFold

Designed at **DeepMind** by Jumper J, et al (2021)

Predicts the 3D structure from the **amino acid sequence**

How does it perform in **domain boundary** prediction?

14th Critical Assessment of Protein Structure Prediction (CASP14)

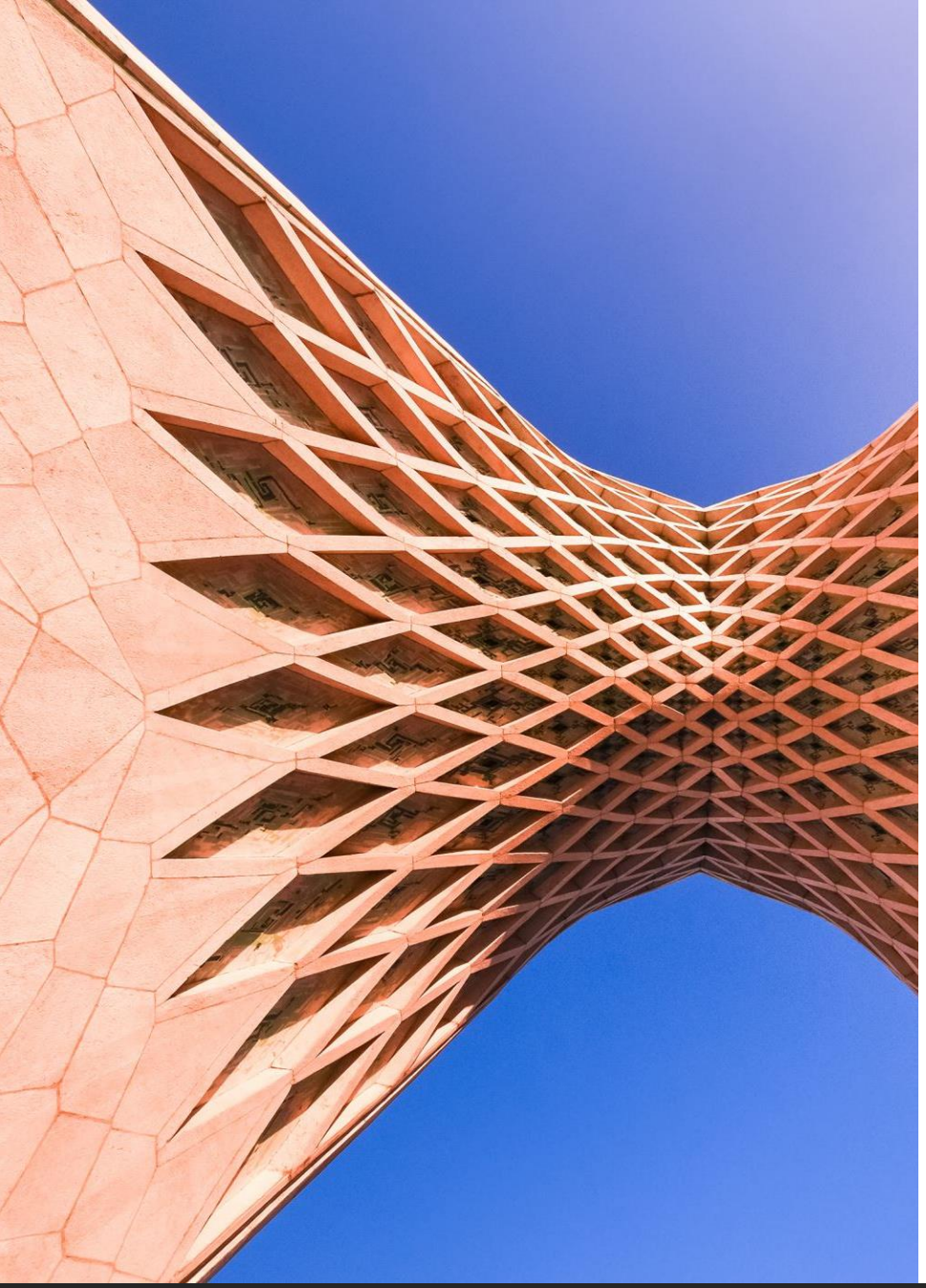


Why do we care about AlphaFold?

Understanding how well it predicts domain boundaries may shine light into whether there is room for the algorithm to **utilise this information** as well for 3D structure prediction

Interested to see if it **compares** with state-of-the-art methods in domain boundary prediction

Problem: How to evaluate **AlphaFold** in domain boundary prediction?



We built a system

Take our dataset and find **AlphaFold** predicted structures of the chains we are using

Validate that the chain in our dataset **matches** exactly the chain in the AlphaFold predicted structure (differences in the databases we collect the data from)

Assign domains to the 3D structure using state-of-the-art domain assignment software using the structure of the protein

Evalute

How does AlphaFold perform?

Results from the evaluation of AlphaFold on a subset of our data (503 chains)

Method	Domain Number Prediction						Boundary prediction
	Single-domain		Multi-domain		All		
	Pre	Rec	Pre	Rec	Acc	MCC	DBD
AlphaFold	**1.0000**	**0.8960**	**0.9483**	**1.0000**	**0.9642**	**0.9217**	0.3173

Conclusion

Questions	Conclusion	Caveat
How do encoding methods compare?	When trained using ESM our model performed better overall	• The difference in the domain boundary distance when using ESM and CARP is not statistically significant - maybe noise • The hyperparameter phase of the model was run using ESM so there may be more optimal values of hyperparameters when using CARP
How does our model perform and compare?	Our method is very good at classifying proteins as single or multi-domain - Better than most state of the art	• Poor performance when predicting the precise position of a boundary
How does AlphaFold perform in domain boundary prediction?	Very high scores	• Dataset could be larger than 503 data points • Performance could show to be even better if a better domain assignment tool was used • Can't compare with state-of-the-art because it was not tested on the same independent dataset

Thank you!